

A Verifiable Search Agent Methodology at Scale for Political Elite Research

Final author list TBD

Abstract

Large- N political-elite datasets increasingly rely on digital traces, but scaling elite attribute coding with “search-enabled LLMs” raises three methodological hazards for inference and replication: (i) *non-verifiability* (answers without durable evidence trails), (ii) *non-comparability* (open-ended labels that drift across countries and time), and (iii) *temporal leakage* (using post-period information to code pre-period attributes, a close analogue of post-treatment conditioning in causal inference). We present a *verifiable search agent* methodology that treats web retrieval as a measurement instrument whose inputs, outputs, and timing are explicitly governed and archived. The agent executes short, provenance-preserving search sessions; constrains decisions to expert-supplied closed codebooks; produces structured, citation-backed outputs; abstains under weak or conflicting evidence; and stores complete traces for audit and re-analysis. In a party-affiliation labeling task spanning 114 countries ($N = 34,618$ matched leader–records; 1,209 party labels), the high-confidence accuracy is 0.860 while expanding usable coverage by 20.8% (12,898 previously missing labels) with a conservative abstention layer.

1 Motivation: why “just use an LLM” is not enough

While large language models possess extensive parametric knowledge, relying solely on their internal weights to code elite political attributes introduces unacceptable risks for scientific research. Standalone LLMs lack inherent temporal awareness, meaning they cannot reliably distinguish between a politician’s affiliation in 2010 versus 2024 without risking anachronistic errors.

Furthermore, their tendency to confidently hallucinate niche factual details—especially for less prominent regional figures—and their fundamental inability to provide verifiable, reproducible primary citations make their raw outputs unsuitable for rigorous quantitative analysis. Consequently, treating an LLM as a static oracle rather than a dynamic reasoning engine inevitably exposes researchers to three fundamental methodological hazards that compromise dataset integrity.

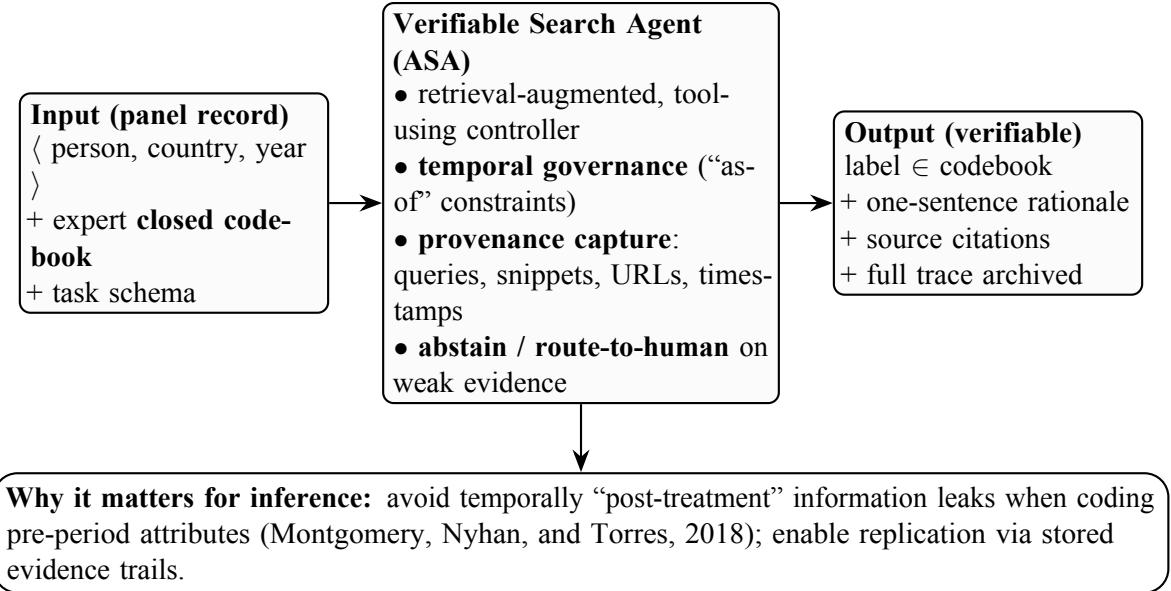


Figure 1: Visual abstract: the ASA treats web retrieval as a governed, auditable measurement process rather than an opaque chat interaction.

Digital sources make it feasible to code political-elite attributes at scale, but research-grade datasets impose requirements that differ from interactive question answering. In practice, a default workflow—“ask a search-enabled model for the answer”—tends to fail on three dimensions central to quantitative political science.

Auditability and replication. Elite attribute labels are *measurements*. When the underlying retrieval context is not archived (queries, results, snippets, URLs, timestamps), a label is difficult to verify and cannot be re-audited when coders disagree or sources change. This is especially problematic for downstream users who need to understand measurement error and may wish to apply alternative inclusion rules (e.g., discarding any label lacking a primary source).

Cross-national comparability. Many elite variables are *categorical* and country-specific (e.g., party labels). “Open-world” generation invites label drift (synonyms, translations, rebrandings), undermining comparability across countries and years. Closed codebooks supplied by domain experts are a natural remedy, but generic search-enabled chat systems do not typically enforce them.

Temporal leakage and post-treatment bias. Political elites switch parties, offices, and coalitions. If coding for year t uses sources written after t (e.g., biographies updated in 2025), the measurement can incorporate information that is itself caused by post- t outcomes. This is a close analogue of conditioning on post-treatment variables in causal inference (Montgomery, Nyhan, and Torres,

2018): a label intended to reflect the pre-period state may be contaminated by later events, inducing bias in estimated relationships. Temporal governance therefore belongs *inside* the measurement protocol, not in ad hoc post-hoc cleaning.

Contributions. We contribute a verifiable search-agent methodology for dataset construction that (i) treats retrieval as a *measurement protocol* with explicit “as-of” constraints to reduce temporal leakage and post-treatment contamination (Kaufman, Rosset, and Perlich, 2012; Montgomery, Nyhan, and Torres, 2018), (ii) enforces expert-supplied *closed codebooks* to preserve cross-national comparability, (iii) produces evidence-first, citation-backed labels while *abstaining* under weak or conflicting evidence, and (iv) archives complete traces (queries, snippets, URLs, timestamps) to enable auditing and replication consistent with calls for reproducible computational research (Peng, 2011; Stodden, Guo, and Ma, 2013).

Table 1: Common hazards in search-enabled coding and ASA design responses.

Hazard	ASA response
Non-verifiability (answers without durable evidence trails)	Evidence-first outputs with explicit citations; full trace capture (queries, snippets, URLs, timestamps) for audit and re-analysis.
Non-comparability (label drift across countries/time)	Closed-world decisions constrained to expert codebooks, plus conservative post-processing that normalizes benign surface variation without introducing new classes.
Temporal leakage (post-period information used for pre-period labels)	Explicit “as-of” rules with date verification where feasible; warnings and abstention rather than silently relying on potentially post-period sources.

2 Methodology: a verifiable search agent

We present a protocol—implemented as the `asa` software stack—that makes agent-based retrieval *verifiable by construction*. The design has four core commitments:

1. **Closed-world decisions.** All predictions must lie in an expert-supplied codebook (e.g., the parties observed in a country-year panel).
2. **Evidence-first outputs.** Each label is paired with a terse rationale anchored to explicit source citations.
3. **Provenance preservation.** The system archives the full interaction trace: queries, tool responses, extracted snippets, URLs, and timestamps.
4. **Conservative abstention.** Under weak or conflicting evidence, the agent abstains (or routes to humans) rather than guessing.

2.1 Task formalization

For each target record i , the inputs are: a person name, a target country, an observation year t_i , and a closed codebook \mathcal{C}_i supplied by expert coders. The agent must return either (a) a label $c \in \mathcal{C}_i$ with citations, or (b) an abstention.

2.2 System design

Operationally, the agent runs a short search session with read-only retrieval tools (e.g., Wikipedia and general web search), compiles candidate evidence, and emits a structured JSON result (label, one-sentence justification with citations, and a confidence category). This design follows retrieval-augmented, tool-using agent patterns in the recent NLP literature (Lewis et al., 2020; Yao et al., 2022). A conservative post-processor then normalizes the label against the codebook: exact-match acceptance first, followed by a guarded fuzzy match for benign surface variation (punctuation, plurals, acronyms) using high similarity thresholds. This reduces typographic drift without introducing new classes.

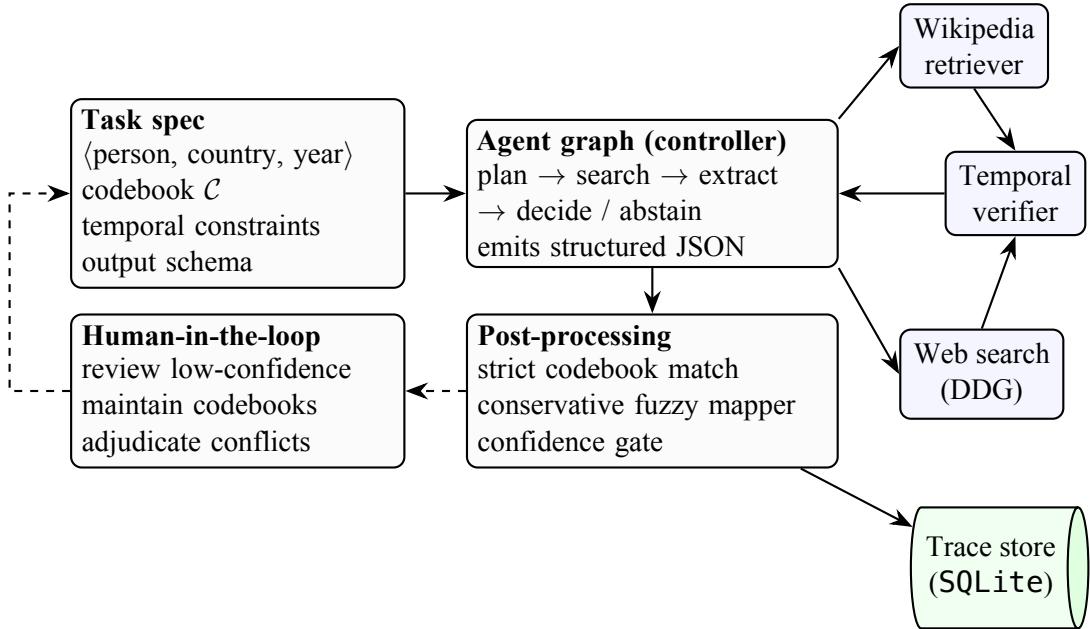


Figure 2: Design visualization: a governed agent graph with temporal verification, conservative normalization, abstention, and an auditable trace store.

2.3 Temporal governance to prevent leakage

The agent treats time as part of the measurement protocol. Each run is parameterized by an “as-of” rule: sources should be published before a cut date (e.g., $t_i + 1$ year) or within a target window

around t_i . When feasible, the system extracts publication dates from retrieved pages and enforces the constraint strictly; otherwise it falls back to best-effort warnings and abstention rather than silently using potentially post-period material. This guards against a common dataset-construction failure mode: coding pre-period attributes with post-period knowledge.

Temporal controls are crucial because many online sources are retrospective and continuously updated: Wikipedia leads, later biographies, and obituaries often summarize whole careers, including events after t_i . If such post- t_i material is used to code a pre-period attribute, the measurement can incorporate information that is itself downstream of later outcomes (e.g., a party switch or coalition realignment that occurs after t_i), creating look-ahead/data leakage and a close analogue of post-treatment conditioning. This can both inflate apparent coding accuracy (by using information unavailable at the time) and bias downstream relationships by contaminating “baseline” covariates with post-period events (Kaufman, Rosset, and Perlich, 2012; Montgomery, Nyhan, and Torres, 2018).

2.4 Provenance preservation and storage

All tool interactions and intermediate artifacts are persisted (including URLs and timestamps) to enable: (i) verification of any individual label by following its citations, (ii) replication of aggregate statistics given the same trace store, and (iii) retrospective audits when codebooks change or new sources appear. The storage design also supports alternative downstream rules (e.g., only keep labels supported by government sources).

3 Case study: party affiliation inference at scale

We evaluate ASA on a verifiable elite attribute: party affiliation. Using the subset of leader–records with expert codings, the agent attains an overall high-confidence accuracy of 0.860 across 114 countries ($N = 34,618$), while expanding coverage by 20.8% (12,898 records) through conservative augmentation. Low-confidence outputs are withheld by design, trading recall for precision; the mean share withheld is 0.250.

4 Discussion: design tradeoffs and practical guidance

Cost and scaling. The methodology is built for high throughput: each record triggers a short, budgeted tool-usage episode and yields a structured output that can be scored, filtered, and aggregated without manual parsing. Using smaller instruction-tuned models for orchestration, caching retrieval

Table 2: Sample agent traces. Text content truncated for readability (and may contain typographical errors as present in native source). Links are clickable. Full traces contain many more sources.

Field	Content
Entry 1: Syleiman Abusaidovich Kerimov — Russian Federation (1999)	
Country	Russian Federation
Year	1999
Person	Syleiman Abusaidovich Kerimov
Wikipedia	Page: Ashot Egiazaryan Summary: Ashot Gevorkovich Egiazaryan (Russian: Ашот Геворкович Егиязарян; Armenian: Աշոտ Եղիազարյան; born...)
Search 1	In the spring of 1998, Yeltsin dismissed Chernomyrdin as head of government and in 1999 Yeltsin's administration backed a newly formed party, Un...
URL 1	https://en.wikipedia.org/...
Search 2	OURHOMEISRUSSIAPARTYOurHomelRussia(Nash Dom—Rossiya, or NDR) was a sociopolitical movement and a ruling party from 1996 to 1998. Source for i...
URL 2	https://www.encyclopedia.com/...
Entry 2: Jasminka Stanojevic — Serbia (2018)	
Country	Serbia
Year	2018
Person	Jasminka Stanojevic
Wikipedia	Page: Supreme Court (Serbia) Summary: The Supreme Court (Serbian: Врховни суд, romanized: Vrhovni sud) is the court of last resort in Serbia...
Search 1	This article lists political parties in Serbia, including parties that existed in the Kingdom of Serbia between the early 1860s and 1918. A kol...
URL 1	https://en.wikipedia.org/...
Search 2	Imali su dve i četiri godine kad smo izbegli iz Knina. Kad bi neko pokucao na vrata, vikali bi: „Tata, tata”. Tri godine nakon progona sazna...
URL 2	https://www.kurir.rs/...
Entry 3: Mihai STROE — Romania (2011)	
Country	Romania
Year	2011
Person	Mihai STROE
Wikipedia	Page: Adrian Stroe Summary: Adrian Stroe (born 24 October 1959), known as The Taxi Driver of Death, is a Romanian serial killer responsible ...
Search 1	Născut în București și cu origini în comuna argeșeană Morărești, fost medaliat cu aur la olimpiada internațională de informatică, Mihai Stroe(...)
URL 1	https://adevarul.ro/...
Search 2	Mihai STROE Parliamentary activity in legislature 2008–2012 DEPUTY Constituency no.38 TULCEA, uninominal college no.2 Member al PDL, deputatul...
URL 2	https://www.cdep.ro/...
Entry 4: Matsie Angelina Motshekga — South Africa (2018)	
Country	South Africa
Year	2018
Person	Matsie Angelina Motshekga
Wikipedia	Page: Angie Motshekga Summary: Matsie Angelina "Angie" Motshekga (born 19 June 1955) is a South African politician and educator who is currently...
Search 1	Matsie Angelina "Angie" Motshekga (born 19 June 1955) is a South African politician and educator who is currently serving as the Minister of Defense...
URL 1	https://en.wikipedia.org/wiki/Angie_Motshekga
Search 2	Motshekga was elected the national president of the African National Congress Women's League (ANCWL) in 2008, defeating the League's secretary-general...
URL 2	https://www.sahistory.org.za/people/matsie-angelina-motshekga-angie-motshekga

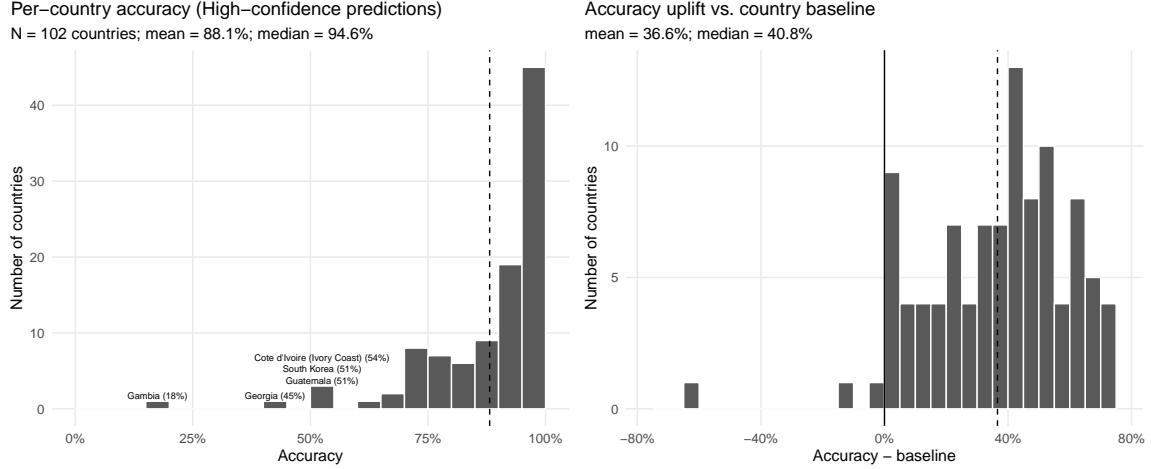


Figure 3: Agent performance in predicting party across countries with sufficient expert overlap. High-confidence predictions are evaluated against expert labels.

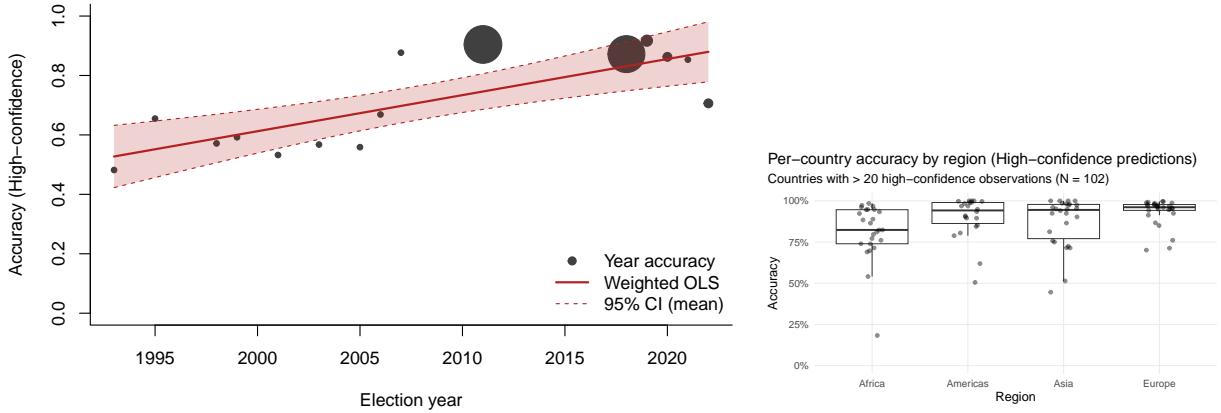


Figure 4: Performance heterogeneity by time (left) and region (right).

outputs, and routing only uncertain cases to humans yields substantial cost savings relative to fully manual coding or repeated interactive browsing.

What this approach does (and does not) guarantee. Provenance capture makes the *evidence* for each label inspectable, but it does not magically eliminate ambiguity in the world. The point of abstention is to keep ambiguity from silently becoming noise. Similarly, temporal governance reduces leakage risk, but cannot fully recover historical truth when the web record is sparse. In practice, auditability means a downstream user can inspect a label by following its recorded citations and can re-apply alternative inclusion rules (e.g., require government sources only) using the archived trace, consistent with calls for reproducible computational work (Peng, 2011; Stodden, Guo, and Ma, 2013). Key failure modes include (i) sources without reliable publication dates, (ii) sparse

Worked example: why “as-of” governance matters. Consider a record with observation year t_i and a label intended to reflect the pre-period state (e.g., party affiliation at t_i). A naive workflow that reads a current biography or Wikipedia lead may inadvertently use retrospective summaries that incorporate events after t_i , creating look-ahead/data leakage (Kaufman, Rosset, and Perlich, 2012) and a close analogue of post-treatment conditioning (Montgomery, Nyhan, and Torres, 2018). ASA instead treats time as part of the measurement protocol: it constrains retrieval to sources published before a cut date (or within a target window), and it abstains when the available evidence cannot be shown to be pre-period. Table 2 illustrates the resulting trace style: multiple sources are retrieved and recorded alongside URLs, allowing a reader to verify what evidence the label relied on under the declared temporal rule.

Figure 5: A compact intuition for temporal governance and trace-based verification.

historical web records for earlier years, and (iii) entities with name ambiguity across languages; ASA responds by warning, tightening evidence requirements, and abstaining rather than imputing.

When to use ASA. ASA is best suited to *verifiable attributes* with authoritative sources (party membership, office holding, election outcomes) and stable codebooks. For sensitive or non-verifiable attributes (e.g., ethnicity without self-identification), we recommend stricter abstention and explicit ethical review.

5 Conclusion

For political-elite research, the central question is not whether LLMs can answer factual questions, but whether they can be integrated into a *measurement protocol* that is auditable, comparable, temporally well-defined, and cost-effective at scale. The verifiable search agent methodology operationalizes these requirements through closed codebooks, evidence-first outputs, strict trace preservation, and conservative abstention. This design turns retrieval-augmented agents into replicable instruments for dataset construction rather than opaque assistants.

References

- Kaufman, Shachar, Saharon Rosset, and Claudia Perlich (2012): “Leakage in data mining: Formulation, detection, and avoidance”. In: *ACM Transactions on Knowledge Discovery from Data*, no. 4, vol. 6, 15:1–15:21. DOI: [10.1145/2382577.2382579](https://doi.org/10.1145/2382577.2382579).
- Lewis, Patrick et al. (2020): “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*.

- Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres (2018): “How conditioning on post-treatment variables can ruin your experiment and what to do about it”. In: *American Journal of Political Science*, no. 3, vol. 62, pp. 760–775.
- Peng, Roger D (2011): “Reproducible research in computational science”. In: *Science*, no. 6060, vol. 334, pp. 1226–1227. DOI: [10.1126/science.1213847](https://doi.org/10.1126/science.1213847).
- Stodden, Victoria, Peixuan Guo, and Zhaokun Ma (2013): “Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals”. In: *PLOS ONE*, no. 6, vol. 8, e67111. DOI: [10.1371/journal.pone.0067111](https://doi.org/10.1371/journal.pone.0067111).
- Yao, Shunyu et al. (2022): *ReAct: Synergizing Reasoning and Acting in Language Models*. arXiv: [2210.03629 \[cs.CL\]](https://arxiv.org/abs/2210.03629).